Cross-Institutional Reuse of a Problem Statement Knowledge Base

Steven H. Brown, M.D.[1,2], Dario A. Giuse, Dr.Ing. [2]

(1) Veterans Affairs Medical Center, Nashville TN.

(2) Division of Biomedical Informatics, Informatics Center, Vanderbilt University Medical Center, Nashville, TN.

This article describes client and server applications for a problem statement knowledge base derived from a large corpus of provider entered terminology. The current status and potential for integration of the server into the Vanderbilt University Medical Center computing environment are discussed. Finally, an experiment in multiple dimensions of reuse for problem list terms is introduced, and possible strategies to mediate between free text and coded data are examined.

INTRODUCTION

The issue of representing clinical knowledge usefully and unambiguously is among the most difficult in medical informatics [1]. Fully coded representations, which are ideal for precise communication and for computer processing, are notoriously expensive to create and may place a heavy burden on already overworked providers. Moreover, the quality of the coding effort itself is questionable [2]. Totally uncontrolled vocabularies, on the other hand, make it easy to record information but shift the load of interpretation and disambiguation to the downstream users of the data.

This article introduces an experiment in multiple dimensions of reuse for problem list terms, and examines possible strategies to mediate between free text and coded data. The terms were derived by synthesizing a large set of provider-entered electronic problem lists from the THERESA[R] system at Grady Memorial Hospital [3]. In this experiment, the authors have attempted to use this rich set of terms as part of the ongoing effort to create an integrated, physician-maintained electronic medical record at Vanderbilt University Medical Center (VUMC). The experiment is aimed at allowing VUMC physicians to begin recording and communicating patients' problem lists immediately, even before the various clinical specialties develop a consensus over which sets of terms should be used to describe patients seen during consults and outpatients visits.

The experiment examines two forms of reuse. The first dimension of reuse is inter-institutional. In principle, terms that were collected at Grady may or may not apply to patients seen at Vanderbilt. However, it is the authors' hypothesis that the size of the data set, and the variety of clinical situations in which the terms were collected, should in fact ensure good coverage for Vanderbilt patients. The experiment will provide a measure of how true that hypothesis holds.

The second dimension is intra-institutional. The VUMC problem lists are currently captured as unstructured free text. However, the structural relationships superimposed on the Grady set of terms could be used to classify and modify the VUMC free-text terms, probably using a semi-automated process. Therefore, even terms that are currently unstructured could be "cleaned up" and converted to coded form with all the attendant benefits. Terms derived from Grady data could potentially be reused both as seeds for the VUMC controlled vocabulary, and to facilitate future structuring of currently free-text terms.

KNOWLEDGE BASE

The problem statement knowledge base was derived from provider-entered electronic problems lists from the THERESA system at Grady Memorial Hospital. Grady Memorial Hospital is a 890 bed inner city public hospital which had 770,651 outpatient visits and 37,304 admissions in 1996. Its Medical Staff, residency training programs, and medical students are affiliated with Emory University or Morehouse University.

The THERESA system is a comprehensive computerized medical record system in place at Grady Memorial Hospital which includes provider entered history and physicals, progress notes, problem lists, and a variety of ancillary data [3]. As of January 1,

1091—8280/97/$5.00 © 1997 AMIA, Inc.

1997, THERESA contained 49,800,000 documents on 1.1 million patients containing 1.4 billion searchable data items and 3.5 billion characters of free text. Provider entered problem lists have been a heavily used core module of THERESA since the module's deployment in December of 1989. The areas with the heaviest use of the problem list application are Medicine and Obstetrics / Gynecology. In the Grady environment problems are defined as "anything that requires management or diagnostic work up; this includes social and demographic problems" [4]. We employ this definition throughout this paper.

The data set used to create the knowledge base consisted of all problem statements entered between December 1989 and May 1995. This totals 891,000 problem statements on 113,000 patients. A variety of automated, manual, and hybrid techniques were applied to the raw dataset to create the problem knowledge base. The goals of this process were to define the minimal set of clinically significant terms to describe the raw data while preserving meaning and capturing interrelationships and frequency data. The process of terminology 'distillation' involved equating redundant or synonymous problems, and splitting expressions containing multiple clinically meaningful concepts. Issues which arose during the processing include misspellings, abbreviations, modifiers, redundancy, synonymy, and lexical normalization.

The problem statement knowledge base consists of several files. The final vocabulary contains 15,553 terms. Each term is assigned a unique serial number and is associated with a frequency computed from the number of occurrences in the initial raw dataset of 891,770 statements. There are 91,125 mappings of clinically equivalent problems, and split multi-concept problems. Files containing weighted pointers between problems currently exist for three types of relationships: a) problems which co-occur on the same line; b) problems expressed as a differential diagnosis; and c) problems expressed as a sign-symptom complex. Finally, a weighted file lists extracted modifiers by frequency of association with each problem. A separate paper describing details of the derivation process and resulting knowledge base is under preparation.

THE VUMC PROBLEM LIST STRATEGY

The work described in this paper is being conducted at Vanderbilt University Medical Center (VUMC). The creation of an integrated application that supports physician-maintained problem lists, allergies, and medications has become an important component of VUMC's IAIMS effort [5] to provide a comprehensive computerized patient record. This effort, which uses VUMC's fast-track approach to building an IAIMS [6], has thus far produced a successful direct provider order entry implementation [7] and a large scale, full-text integrated patient record [8]. Because the order entry system does not yet cover outpatients, the chosen strategy has been to develop modular, easy-to-use components that allow problems, allergies, and medications to be recorded electronically.

During the design of these components, clinical users expressed two requirements. First, the application(s) should be very easy to use and should not disrupt the busy schedule of clinical care. Second, it should be possible to start using the tool immediately, well before completion of the lengthy consensus-building processes that will lead to the adoption of terminology standards for the entire institution.

Problem list terms being collected at VUMC come from three main sources: a Web-based problem list application; the order-entry front end; and dictated clinic notes. All three sources currently allow providers to enter problems as free text

These requirements strongly influenced the work described in this paper. The knowledge base produced from large store of problem list terms that had been entered by care providers at Grady is a sizable initial set of terms that can be used to map and examine terms entered at Vanderbilt. Even more importantly, the extensive work on creating structural relationships among terms (and components of terms) allows newly entered problems to be automatically classified within the same structures, potentially giving physicians suggestions about other potentially associated problems.

KNOWLEDGE BASE SERVER

To explore the reusability of the Grady terms, the authors have developed a knowledge base server which allows providers to enter individual problem terms and explore their relationship to other terms. The problem statement knowledge base server is a generic, easily accessible resource which features TCP/IP sockets based connectivity, simply formatted ASCII output, and multiprocessing capabilities. Sockets represent a widely available technology which is the standard method of communication between applications in our environment. ASCII encoded, minimally formatted textual output from the server simplifies both client and server side processing. Complex output formatting would require additional server computation, and would make client side recasting of data more complicated.

The server is written in Perl, and runs in the Unix environment. Each connection from a client forks a child process which services the request. At start-up time the server reads raw indices and data files of the problem statement knowledge base and computes compiled indices. The compiled indices subsequently remain in memory while the server is running. This approach allows the initial processing of the data to occur only once, and allows requests from clients to be handled very efficiently.

The server accepts two types of requests. The first is a request to match an input string against the problem terminology. The second is a request to examine all links for a specific problem statement. Terminology look-ups permit multi-key searches for text fragments. A scored sorted list is returned to the client application. Sorting occurs on three criteria in the following order: 1) the matching score; 2) the overall frequency of matched term in the knowledge base; and 3) alphabetically. The unique serial number of each textual problem statement is also output by the server. A sample of server output for terminology lookup is in Figure 1.

A request to examine links for a specific vocabulary term is sent by sending the term's unique serial number to the server. The server responds by computing all family members of the term as the transitive closure across all generations of the vocabulary (recall that processing the raw data was a multi-step process). All modifiers associated with the term, and all related problems for the entire family of problems, are reported to the client. An example of server output of associated links for a problem is in figure 2.

Two sample clients are available to access the server. A command line based client opens a socket to the waiting server and transmits and receives data. The user inputs a string or a serial number, and is returned data as described above. This client was used by the authors for early testing of the server.

A second Common Gateway Interface (CGI) based Web client allows user interaction using hypertext and forms. The user enters the first query request by typing a term into an HTML form. Subsequent requests may be entered in the same fashion or by clicking on results of previous queries. The latter capability allows rapid traversal of the knowledge base. The Web client is shown in figure 3.

The CGI Web client is being integrated with the VUMC problem list application, which is also a Web-based program. This will allow VUMC users to submit terms to the server and explore potential mappings to terms and modifiers from the Grady set.

The subsequent stage will explore the use of the server in a more automated way, either as a background activity while the user is entering terms or as an after-the-fact verification and cleanup process.

DISCUSSION

There are both immediate improvements and long term goals for further development of the problem statement knowledge base. Two new types of links are partially developed at the time of writing. The first link type

```
Candidate terms and scores - Top 13 of 245

Score Freq   Problem                                    Serial No.
147   44     MYOCARDIAL INFARCTION ANTERIOR               0016441
147   12     MYOCARDIAL INFARCTION ANTEROSEPTAL           0011691
147    5     MYOCARDIAL INFARCTION ANTEROLATERAL          0041211
147    2     MYOCARDIAL INFARCTION ANTEROAPICAL           0021826
146    8     MYOCARDIAL INFARCTION ANTERIOR ACUTE         1000309
146    5     MYOCARDIAL INFARCTION ANTERIOR SILENT        0007367
146    5     MYOCARDIAL INFARCTION ANTEROSEPTAL ACUTE     0008864
146    4     MYOCARDIAL INFARCTION ANTEROSEPTAL MASSIVE   0023935
146    2     MYOCARDIAL INFARCTION ANTERIOR SUBENDOCARDIAL 0038427
146    2     MYOCARDIAL INFARCTION ANTEROLATERAL ACUTE    0034114
144    4     MYOCARDIAL INFARCTION ANTERIOR NON Q WAVE    0010531
 68 2502     MYOCARDIAL INFARCTION                        1018028
 67 1921     MYOCARDIAL INFARCTION ACUTE                  1000349
```

Figure 1: Sample server output for the input string "myocar infar ant"

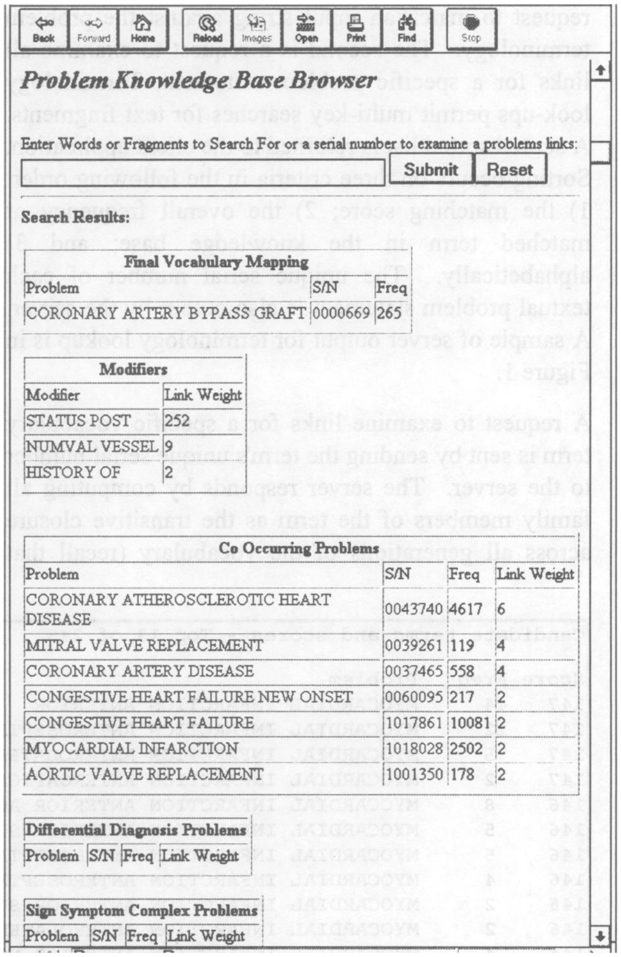
```
The Final Problem Statements are:
MYOCARDIAL INFARCTION ACUTE          1000349     1921

Modifiers are:
HISTORY OF                           716
RULE OUT                             224
STATUS POST                          16
HISTORY                              8
SECONDARY                            4
POSSIBLE                             3
PRESENT                              2
PROBABLE                             2

Co Occurring Problems are:
ANGINA UNSTABLE                      1018033     4432        3
CHEST PAIN                           1017850     6567        2
CARDIAC ARREST                       1011300     227         2
LUNG RUL                             0022064     4           1
CHEST PAIN TIMING PROLONGED          0004881     9           1
CONGESTIVE HEART FAILURE             1017861     10081       1
CHEST PAIN ATYPICAL                  0060117     1502        1
HYPOTENSION                          1014734     391         1

DDX Problems are:
ANGINA UNSTABLE                      1018033     4432        5
```

Figure 2: Sample server output for a request to examine links for specific problem "1000349". Column 2 is the serial number, Column 3 the overall frequency in the knowledge base, and Column 4 the link weight.

associates main and sub-problems from the original dataset. The second type associates problem statements with demographic profiles of the original patients. Indices to support problem lookup against an unprocessed version of the vocabulary will be built. Combining this with the server's existing capabilities to map problem statements forward to final vocabulary promises a novel approach to automated coding. We hypothesize that this approach may permit entry of typical clinical free text, and provide the user with options for final codes that include such niceties as automated correction of misspellings and expansion of abbreviations.

Other goals include development of tools to maintain and evolve the current knowledge base. This includes the addition of new terminology, as well as updating frequency data and link weights. The prospect of a cross institutional component presents additional opportunities. While the Grady terms themselves should prove applicable to Vanderbilt, it seems certain that the relative frequency weights will need to be modified because of the differences between the two patient populations. The comparison of frequencies computed from the Vanderbilt set with those from the original Grady set should provide interesting insights about the relative stability of the frequency weights.

To support the expect request load, the server needs to undergo capacity testing based on estimates of demand within VUMC. Options to improve performance include running multiple servers on different machines, compiling the Perl code, or rewriting the more critical portions of the server in C or C++.



Problem Knowledge Base Browser

Enter Words or Fragments to Search For or a serial number to examine a problems links:

Search Results:

Final Vocabulary Mapping		
Problem	S/N	Freq
CORONARY ARTERY BYPASS GRAFT	0000669	265

Modifiers	
Modifier	Link Weight
STATUS POST	252
NUMVAL VESSEL	9
HISTORY OF	2

Co Occurring Problems			
Problem	S/N	Freq	Link Weight
CORONARY ATHEROSCLEROTIC HEART DISEASE	0043740	4617	6
MITRAL VALVE REPLACEMENT	0039261	119	4
CORONARY ARTERY DISEASE	0037465	558	4
CONGESTIVE HEART FAILURE NEW ONSET	0060095	217	2
CONGESTIVE HEART FAILURE	1017861	10081	2
MYOCARDIAL INFARCTION	1018028	2502	2
AORTIC VALVE REPLACEMENT	1001350	178	2

Differential Diagnosis Problems			
Problem	S/N	Freq	Link Weight

Sign Symptom Complex Problems			
Problem	S/N	Freq	Link Weight

Figure 3: screen image from the CGI-based Web client.

In addition to the immediate benefits derived from integrating the knowledge server with the VUMC problem list application, the work described in this paper will provide an evaluation of the feasibility of reusing a knowledge resource such as the Grady problem statement terms. Considerable effort went into creating, collecting, and then processing the terms to extract the current knowledge base. If the resource can be shown to facilitate the structuring of VUMC problem lists, and more generally the evolution of the VUMC problem list strategy, this will constitute a useful example of reusing a knowledge resource across institutions. Moreover, this work could potentially provide benefits in the opposite direction, because the VUMC data could provide a validation or suggest enhancements to the structure created from the Grady data. This validation could be strengthened even further if data from problem lists created at Grady over the past two years (which are not included in the set used in this study) could be analyzed, yielding a three-way comparison across institutions and across time.

Acknowledgments

This work was supported in part by grants 1 F38 LM00021-01A1 and 1 R55 LM05779-01 from the National Library of Medicine.

THERESA is a product of Medical Systems Development Corporation, Atlanta GA.

References

1. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. JAMIA. 1994;1:35-50.

2. Lloyd SS, Rissing JP. Physician and coding errors in patient records. JAMA. 254(10):1330-6, 1985 Sep 13.

3. Camp HN. Technical challenges, past and future, in implementing THERESA - a one million patient, one billion item computer-based patient record and decision support system. Proc Health Care Information Infrastructure. SPIE, Bellingham, WA, 1995.

4. Weed LL. Medical Records, Medical Education, and Patient Care: The problem-oriented Record as a Basic Tool. Year Book Medical Publishers, Inc, Chicago 1969.

5. Stead WW, Baker W, Harris TR, Hodges TM, Sittig DF. A fast track to IAIMS: the Vanderbilt University strategy. Proc 16th SCAMC, Baltimore, MD, 1992;527-31.

6. Stead WW, Borden R, Bourne J, Giuse DA, et al. The Vanderbilt University Fast Track to IAIMS: Transition from Planning to Implementation. JAMIA. 1996;3:308-317.

7. Geissbuehler A., Miller RA. A New Approach to the Implementation of Direct Care-Provider Order Entry. AMIA Fall Symposium, Washington, DC, October. 1996; 689-693.

8. Giuse DA, Mickish A. Increasing the Availability of the Computerized Patient Record. AMIA Fall Symposium, Washington, DC, October. 1996; 633-637.